D766–D773 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 24 October 2018*
doi: 10.1093/nar/gky955

GENCODE reference annotation for the human and mouse genomes

Adam Frankish[1], Mark Diekhans[2], Anne-Maud Ferreira[3], Rory Johnson[4,5], Irwin Jungreis[6,7], Jane Loveland[1], Jonathan M. Mudge[1], Cristina Sisu[8,9], James Wright[10], Joel Armstrong[2], If Barnes[1], Andrew Berry[1], Alexandra Bignell[1], Silvia Carbonell Sala[11], Jacqueline Chrast[3], Fiona Cunningham[1], Tomás Di Domenico[12], Sarah Donaldson[1], Ian T. Fiddes[2], Carlos García Girón[1], Jose Manuel Gonzalez[1], Tiago Grego[1], Matthew Hardy[1], Thibaut Hourlier[1], Toby Hunt[1], Osagie G. Izuogu[1], Julien Lagarde[11], Fergal J. Martin[1], Laura Martínez[12], Shamika Mohanan[1], Paul Muir[13,14], Fabio C.P. Navarro[8], Anne Parker[1], Baikang Pei[8], Fernando Pozo[12], Magali Ruffier[1], Bianca M. Schmitt[1], Eloise Stapleton[1], Marie-Marthe Suner[1], Irina Sycheva[1], Barbara Uszczynska-Ratajczak[15], Jinuri Xu[8], Andrew Yates[1], Daniel Zerbino[1], Yan Zhang[8,16], Bronwen Aken[1], Jyoti S. Choudhary[10], Mark Gerstein[8,17,18], Roderic Guigó[11,19], Tim J.P. Hubbard[20], Manolis Kellis[6,7], Benedict Paten[2], Alexandre Reymond[3], Michael L. Tress[12] and Paul Flicek[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA, [3]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland, [4]Department of Medical Oncology, Inselspital, University Hospital, University of Bern, Bern, Switzerland, [5]Department of Biomedical Research (DBMR), University of Bern, Bern, Switzerland, [6]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vasser St, Cambridge, MA 02139, USA, [7]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA, [8]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [9]Department of Bioscience, Brunel University London, Uxbridge UB8 3PH, UK, [10]Functional Proteomics, Division of Cancer Biology, Institute of Cancer Research, 123 Old Brompton Road, London SW7 3RP, UK, [11]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, Barcelona, E-08003 Catalonia, Spain, [12]Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, [13]Department of Molecular, Cellular & Developmental Biology, Yale University, New Haven, CT 06520, USA, [14]Systems Biology Institute, Yale University, West Haven, CT 06516, USA, [15]Centre of New Technologies, University of Warsaw, Warsaw, Poland, [16]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA, [17]Program in Computational Biology & Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA, [18]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA, [19]Universitat Pompeu Fabra (UPF), Barcelona, E-08003 Catalonia, Spain and [20]Department of Medical and Molecular Genetics, King's College London, Guys Hospital, Great Maze Pond, London SE1 9RT, UK

Received August 15, 2018; Revised September 20, 2018; Editorial Decision October 02, 2018; Accepted October 08, 2018

ABSTRACT

The accurate identification and description of the genes in the human and mouse genomes is a fundamental requirement for high quality analysis of data informing both genome biology and clinical genomics. Over the last 15 years, the GENCODE consortium has been producing reference quality gene annotations to provide this foundational resource. The GENCODE consortium includes both experimental and computational biology groups who work to-

*To whom correspondence should be addressed. Tel: +44 1223 492581; Fax: +44 1223 494494; Email: flicek@ebi.ac.uk

gether to improve and extend the GENCODE gene annotation. Specifically, we generate primary data, create bioinformatics tools and provide analysis to support the work of expert manual gene annotators and automated gene annotation pipelines. In addition, manual and computational annotation workflows use any and all publicly available data and analysis, along with the research literature to identify and characterise gene loci to the highest standard. GENCODE gene annotations are accessible via the Ensembl and UCSC Genome Browsers, the Ensembl FTP site, Ensembl Biomart, Ensembl Perl and REST APIs as well as https://www.gencodegenes.org.

INTRODUCTION

The GENCODE consortium produces foundational reference genome annotation for the human and mouse genomes as well as tools and data to maintain and improve these annotations. Our overall goal is to identify and classify, with high accuracy, all gene features in the human and mouse genomes based on defined biological evidence and to make these annotations freely available for the benefit of biomedical research and genome interpretation.

The GENCODE project was founded in 2003 as part of the pilot phase of the ENCODE project to provide reference quality manual gene annotation for the 30Mb (∼1%) of the reference human genome targeted by the ENCODE pilot (1–3). In 2007, we expanded our scope to the whole human genome as the ENCODE project did the same (4,5). In 2012, we began annotating the mouse reference genome to the same standards as human, while continuing to improve the existing gene annotation in both species via targeted reinvestigation of loci flagged by external users and internal QC pipelines. Today, the GENCODE consortium is a long-running partnership of manual annotation, computational biology and experimental groups including four of the founding groups (HAVANA, CRG, Yale and UCSC) and three groups that joined in 2007 (Ensembl, MIT and CNIO).

Our gene annotations are regularly released as the Ensembl/GENCODE gene sets. The gene sets are comprehensive and include protein-coding and non-coding loci including alternatively spliced isoforms and pseudogenes. To produce the annotations, we leverage computational and experimental methods to identify new genes and new transcript isoforms, directing manual annotation to regions requiring expert investigation. The Ensembl/GENCODE annotations are the default human and mouse annotation for the Ensembl project (6), while the UCSC Genome Browser (7) uses the human annotation as default and the mouse annotation as a secondary resource until the mouse clone-by-clone annotation is complete (see below). For each versioned release, the underlying genome annotation is exactly the same whether it is accessed at Ensembl, UCSC or https://genecodegenes.org, although there are minor differences in presentation associated with genome assembly patches and representation of the pseudoautosomal regions on the X and Y chromosomes. We also provide subsets of the annotation as described below. For simplicity, we will here refer to the annotation holistically as GENCODE.

GENCODE is the reference annotation of choice adopted by many large international consortia including ENCODE, GTEx (8), the International Cancer Genome Consortium (ICGC) (9), component projects of the International Human Epigenome Consortium (10), the 1000 Genomes Project, (11) the Exome Aggregation Consortium (EXAC) and Genome Aggregation Database (gnomAD) (12) and the Human Cell Atlas (HCA) (13).

GENCODE ANNOTATION METHODS AND RESULTS

The GENCODE consortium annotates protein-coding genes, pseudogenes, long non-coding RNAs (lncRNAs) and small non-coding RNAs (sncRNAs). We define protein-coding genes as loci where the weight of available evidence supports the presence of a coding sequence (CDS). Evidence for a CDS may come from high-throughput experimental assays, the demonstration of physiological function in the research literature, the observation of homology to a known protein-coding gene, or the interpretation of evolutionary conservation data. Pseudogenes are sequences derived from protein-coding genes, containing disabling mutations such as in-frame stop codons, frameshifting indels, truncations or insertions, or for which there is no evidence of transcription. lncRNA genes are identified by a combination of transcriptional evidence and a lack of potential to be assigned as protein-coding. We do not absolutely require lncRNA genes to be longer than 200 bp, but very few annotated lncRNAs fall below this threshold, as we also require annotated lncRNAs to be free of secondary structures found in known functional sncRNAs. Currently, sncRNAs are almost entirely annotated by computational pipelines that use homology to known sncRNA sequences and predicted secondary structure to identify functional copies.

Our annotation processes use primary transcript and proteomics data, evolutionary conservation, computational methods and curated public databases such as UniProt (14). These data are integrated using a combination of expert manual annotators and computational methods to identify regions of the genome with genic potential, annotate the exon-intron structures of transcripts identified at the locus under investigation and assign a functional classification to both the individual transcript and the locus.

Broad functional classes (referred to as 'biotypes') of protein-coding, pseudogene, lncRNA and sncRNA are assigned as described above. More detailed functional categories are also added. For example, at the locus level we describe the provenance of pseudogenes as processed (derived via retrotransposition), unprocessed (defined by a genome duplication event) or unitary (arising from the lineage specific disruption of an ancestral protein-coding gene). At the transcript level we define transcripts belonging to protein-coding loci as protein-coding, nonsense mediated decay (NMD) (containing a premature stop codon believed likely to lead to the transcript being targeted by the nonsense-mediated decay pathway) or retained intron (containing sequence that is intronic in other transcripts from the locus). Following the structural and functional classification of transcripts, a subset of GENCODE annotation is sub-

ject to targeted experimental validation as described below to ensure consistent high quality of the gene annotation.

To cater for a variety of use cases, we create a number of annotation sets. Examples of these are our 'GENCODE comprehensive' and 'GENCODE basic' gene sets. GENCODE comprehensive includes the complete set of annotations including partial transcripts (i.e. transcripts that are not full length, but represent a unique splice form based on available evidence) and biotypes such as NMD. GENCODE basic is a subset of GENCODE comprehensive that contains only transcripts with full-length CDS. For non-coding loci, GENCODE basic includes the smallest number of transcripts that cover 80% of the exonic features, while ensuring all loci are represented by at least 1 transcript. Computational methods add additional information. For example, APPRIS, described in more detail below, identifies the most likely functional translations at protein-coding loci and TSL (transcript support level) calculates the amount and quality of supporting evidence for each transcript.

Manual annotation

The GENCODE gene set is created by merging the results of manual and computational gene annotation methods. Manual gene annotation has two major modes of operation: clone-by-clone and targeted annotation. 'Clone-by-clone' annotation involves 'walking' across a genomic region, investigating the sequence, aligned expression data and computational predictions for each BAC clone. In doing so, an expert annotator investigates all possible genic features and considers all possible annotations and biotypes simultaneously. We believe this approach carries substantial advantages. For example, the decision to annotate a locus as protein-coding or pseudogenic benefits from being able to weigh both possibilities in light of all available evidence. This process helps prevent false positive and false negative misclassifications. Targeted annotation is designed to answer specific questions such as 'is there an unannotated protein-coding gene in this position?' Ranked target lists are generated by computational analysis based, for example, on transcriptomic data, shotgun proteomic data or conservation measures. Over the last two years mouse annotation has been dominated by the clone-by-clone approach while the human genome has been refined entirely via targeted reannotation except for the annotation of human assembly patches and haplotypes released by the Genome Reference Consortium (15), which take a clone-by-clone approach.

Over the last two years, we have focused on two broad areas: completing the first pass manual annotation across the entire mouse reference genome and a dedicated effort to improve the annotation of protein-coding genes in human and mouse.

We have completed the annotation of novel protein-coding genes, lncRNAs and pseudogenes, plus QC and updating previous annotation where necessary for mouse chromosomes 9, 10, 11, 12, 13, 14, 15, 16 and 17. These updates bring the fraction of the mouse genome with completed first pass manual annotation to approximately 97%. In addition, we have continued to work with the NCBI and Mouse Genome Informatics project at the Jackson Laboratory to resolve annotation differences for protein-coding, pseudogene and lncRNA loci. For protein-coding genes this is under the umbrella of the Consensus Coding Sequence (CCDS) project (16).

We have also manually investigated unannotated regions of high protein-coding potential identified by whole genome analysis using PhyloCSF (17) (a tool described in more detail below). In human, this led to the addition of 144 novel protein-coding genes and 271 pseudogenes (of which 42 were unitary pseudogenes). In mouse, we annotated orthologous loci for all but 11 of the 144 human protein-coding genes. We have also revisited the annotation of all olfactory receptor loci in both human and mouse, using RNAseq data to define 5′ and 3′ UTR sequences for ∼1400 loci. In human we have also targeted a 'deep dive' manual reannotation of genes on clinical panels for paediatric neurological disorders to identify missing functional alternative splicing. Incorporating second and third generation transcriptomic data, we reannotated ∼190 genes and added more than 3600 alternatively spliced transcripts, including ∼1400 entirely novel exons and an additional ∼30kb of CDS. We have also completed an effort to capture all recently described unannotated microexons (18) into GENCODE, and further added an additional 146 novel microexons mined from public SLRseq data (19).

As part of the CCDS collaboration with RefSeq, we have checked a large subset of human loci where there was disagreement over gene biotype. Similarly, we have checked all UniProt manually annotated and reviewed (i.e. Swiss-Prot) accessions that lack an equivalent in GENCODE. As a result, we added 32 novel protein-coding loci to GENCODE and rejected more than 200 putative coding loci. Finally, we are manually reviewing genes previously annotated as protein-coding, but with weak or no support based on a method incorporating UniProt, APPRIS, PhyloCSF, Ensembl comparative genomics, RNA-seq, mass spectrometry and variation data (20,21). Of the 821 loci investigated to date, 54 have had their coding status removed while a further 110 potentially dubious cases remain under review.

The approach taken reflects in the kinds of updates captured in the annotation. For example, the targeted reannotation in human leads to the annotation of few novel protein-coding loci but many novel transcripts at updated protein-coding and lncRNA loci. Conversely, in mouse the emphasis on clone-by-clone annotation identifies many more novel loci and transcripts across a broader range of biotypes (Figure 1).

Computational annotation of small RNAs

We annotate small non-coding RNAs (sncRNAs) using a variety of mechanisms. Specifically, miRNA annotations are imported directly from miRBase (22), while tRNAs are identified *ab initio* using tRNAScan-SE (23) although they are not included directly in the gene set. For other classes of sncRNA, including small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and small Cajal body-specific RNAs (scaRNAs), we use a homology-based, computational pipeline (24), which first compares sequences of known RNA families in Rfam (25) to the genome using BLAST (26). This initial step reduces the genomic search space and excludes sequences with sub-optimal alignments

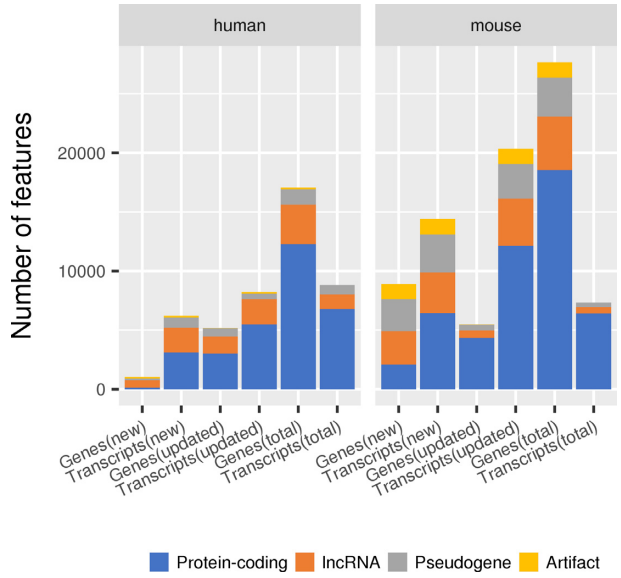

Figure 1. New and updated manually annotated genes and transcripts from July 2016 to June 2018. For both human (left) and mouse (right) the numbers of completely new genes and transcripts, updated genes and transcripts and the total number of manually added or edited genes and transcripts for each of four broad categories of annotation. A new gene annotation can represent a completely *de novo* locus with no overlap with pre-existing annotation or the reclassification of an existing complex locus into multiple loci to better represent the biology of the locus inferred from transcriptomic and/or proteomic data. A new transcript represents the annotation of a unique exon-intron structure, including novel alternative splicing at an annotated locus. Updated genes and transcripts represent pre-existing loci or transcript models that have been edited to improve the representation of biotype (e.g. changed from lncRNA to protein-coding) or structure (e.g. by extension, addition of novel exons).

to the genome. We define putative sncRNA models after clustering top BLAST hits and evaluating these predictions by performing sequence and structure searches against covariance models in the Infernal suite of tools (27).

Pseudogenes

Pseudogene annotations across 18 mouse strains were generated using a combination of manual annotation liftover and computational methods. Additionally, we were able to annotate 88 new human and 131 new mouse unitary pseudogenes relative to each other. Amongst the strains we find roughly 20 unitary pseudogenes per strain. We identified nearly 3000 ancestral pseudogenes conserved across all strains. Meanwhile, ∼20% of the pseudogenes in each strain are strain specific. In line with previous results in human, 15% of pseudogenes exhibit transcriptional activity (bioRxiv: https://doi.org/10.1101/386656).

EXPERIMENTAL ANNOTATION APPROACHES

lncRNA annotation using capture long Seq

Determining the precise boundaries and the exonic structure of low abundant transcripts, such as lncRNAs is challenging. We previously showed that 3′ and 5′ boundaries of lncRNAs annotated in GENCODE V7 (April 2011) were less supported by CAGE and PET tags than those of protein-coding genes, even when accounting for differences in expression (28). Methods to assemble transcript sequences from short sequence reads have also been shown to produce poor results when used to resolve the exonic structure of lncRNAs (29,30). To improve lncRNA annotation, we developed the RNA Capture Long Seq (CLS) method (31). Here, probes are first designed against targeted lncRNAs (or suspected, unannotated lncRNA loci). Full-length cDNAs generated from diverse cell types were captured, resulting in cDNA libraries that are highly enriched for the targeted lncRNAs. Libraries were then sequenced using long-read sequencing technologies (31,32). Our initial efforts created a comprehensive capture library targeting the set of intergenic GENCODE lncRNAs in human and mouse, and used it in a set of matched human and mouse tissues (31). This resulted in novel lncRNA transcripts at 3574 loci in human, and 561 in mouse. The long length of the transcript sequences obtained, often correspond to complete 5′-to-3′ RNA molecules, substantially informed manual annotation. Indeed, CLS produces near manual-quality full-length transcript models at high-throughput scales (32). Our current efforts are to include samples across a more diverse panel of tissues such as fetal timepoints.

Proteomics

Proteomic mass spectrometry datasets are a powerful resource contributing to the validation and annotation of protein-coding genes and transcripts. In GENCODE, we use proteomics data as an additional layer of evidence when defining the structure and protein-coding potential of a genomic locus. We apply strict criteria to the peptide evidences we consider from mass spectrometry datasets (33–35) to minimize the incorporation of false positive and ambiguous or variant peptide species. In highly curated genomes such as human, the contribution from mass spectrometry experiments requires considerable scale of data and effort, with correspondingly small returns. Our experimental efforts in GENCODE incorporate targeted proteomics experiments, specific experimental designs and synthetically generated peptides to find these elusive protein-coding genes.

Annotation validation and RACEseq

We used RT-PCR amplification followed by highly multiplexed sequencing readout (36) to assess the quality of the annotations. This method evaluates low confidence transcribed loci (novel or putative). Splice site loci were systematically experimentally tested in eight tissues (brain, heart, kidney, liver, lung, spleen, skeletal muscle, and testis) by RT-PCR-seq (36). From human GENCODE versions 3 to 19, a total of 18 132 splice junctions were analyzed and experimentally tested. Seventy eight percent of all assessed junctions were confirmed through experimental validation. Similar to the human annotation, we assessed the quality of the mouse annotation. A total of 3956 splice junctions from GENCODE versions M2 and M4 were tested with a validation rate of 53%. Finally, to assess the completeness of the annotations we amplified and sequenced the transcripts of 527 deeply annotated human protein-coding genes, which are routinely used for diagnostic tests by the

UK Genetic Testing Network (UKGTN). We performed 5′- and 3′- nested- RACEs in seven different tissues (brain, testis, heart, kidney, liver, lung, and spleen) followed by long-read sequencing, which revealed 10 380 novel splice junction candidates.

GENCODE ANNOTATION TOOLS

Comparative annotation toolkit

We developed the Comparative Annotation Toolkit (CAT) (37) to leverage the GENCODE annotations of mouse and human to annotate laboratory mouse strains (38) and great apes (39,40). CAT uses whole genome alignments from Cactus (41) to project GENCODE annotations from mouse or human to related species, and then performs a variety of filtering and clean-up steps to generate a high quality annotation set for these other genomes. The GENCODE M11 mouse annotation was used with CAT to annotate 16 laboratory mouse strains, and these annotations are available in Ensembl. Over 20 000 protein-coding and 12 000 non-coding genes were comparatively annotated in each lab strain. Novel gene predictions using Comparative Augustus (42) also found an average of 22 new loci in classical strains, including the discovery of the gene *Efcab3-like* in the reference mouse, which was included in subsequent GENCODE releases. Additionally, the GENCODE 27 (August 2017) human annotation set was used to annotate chimpanzee, gorilla and orangutan, and these annotations were incorporated into Genbank, with over 19 000 protein-coding and 36,000 non-coding genes comparatively annotated in all of the great apes.

APPRIS

The APPRIS Database (http://appris-tools.org) (43) was developed to provide annotations for alternative splice variants. APPRIS also determines principal splice isoforms based on cross-species conservation and the conservation of protein structure and function. Most coding genes have a single dominant protein isoform and this main isoform is almost always the APPRIS principal isoform (44).

APPRIS maintains up-to-date annotations for the GENCODE and RefSeq reference sets and has been extended to the UniProtKB proteome and to six model species as well as human and mouse (45). Technical improvements include incremental improvements to the core modules that make up the APPRIS pipeline, the implementation of a UCSC Track Hub to make annotation access easier, and Docker images to allow the execution of the annotation pipeline (45).

APPRIS is an integral part of the pipeline for the prediction of potential non-coding genes (20). For the GENCODE 27 (August 2017) human annotation the completed pipeline flagged 2432 genes.

PhyloCSF

Comparative genomics is one of the most powerful tools available for distinguishing protein-coding genomic regions. Previously, we developed PhyloCSF to support annotation of coding sequences based on the alignment of multiple genome sequences (17). As described above, we combine whole-genome PhyloCSF data with experimental evidence and expert manual annotation to detect novel coding sequences. The workflow begins with PhyloCSF scores computed on every codon in the human genome in each of the six reading frames; applies a Hidden Markov Model to these scores to find candidate coding intervals; excludes intervals previously annotated as coding or pseudogene, or antisense to such intervals, as well as very short intervals; and uses a Support Vector Machine to prioritize the resulting 'Novel PhyloCSF Regions'. We have created publicly available PhyloCSF track hubs for viewing the whole-genome PhyloCSF data and novel PhyloCSF Regions from human and mouse in the UCSC and Ensembl genome browsers.

Pseudopipe

Pseudopipe identifies and annotates pseudogenes across the genome (46). It takes as input an organism's protein-coding gene set and searches for homology across the genome using BLAST. Hits overlapping functional genes are removed and the remaining hits are then assembled into pseudogene annotations. Each annotation is also assigned a parent gene, the functional paralog that gave rise to the pseudogene, as well as a biotype (processed, duplicated, or ambiguous). Unitary pseudogenes are also identified via Pseudopipe by using a different organism's protein-coding gene set as the input. We inform our annotation with results from Retrofinder (47) and RCPedia (48). In addition to our core annotation files, further information is available at http://www.pseudogene.org. These computational annotations are then combined with manual annotations in order to produce the full pseudogene complement. Pseudogene annotations are given a confidence level based on the intersection with manual annotations. Annotations detected by both the computational pipelines and manual annotators are assigned level 1, those only detected by manual annotators are given level 2, and the consensus annotations detected by PseudoPipe and RetroFinder are given level 3 and made available in a separate annotation file at https://www.gencodegenes.org.

DATA ACCESS

Versioned GENCODE gene sets are currently released approximately four times a year for mouse and twice a year for human. This asymmetric update pattern reflects the fact that the first pass of the human annotation was completed in GENCODE 15 (January 2013), while the mouse first pass is approaching completion (expected for GENCODE M20) and therefore has been the subject of more intensive annotation. The most recent release of the human geneset is GENCODE 29 (October 2018), while the most recent mouse update is GENCODE M19 (October 2018). Each release incorporates the continuous updates arising from expert manual annotation. Figure 2 shows the increase in the numbers of genes and transcripts in human and mouse GENCODE releases over the past two years. The human genesets look relatively static, although headline figures do not capture updates made to existing annotation and the balancing effect of both adding and removing loci during a release cycle.

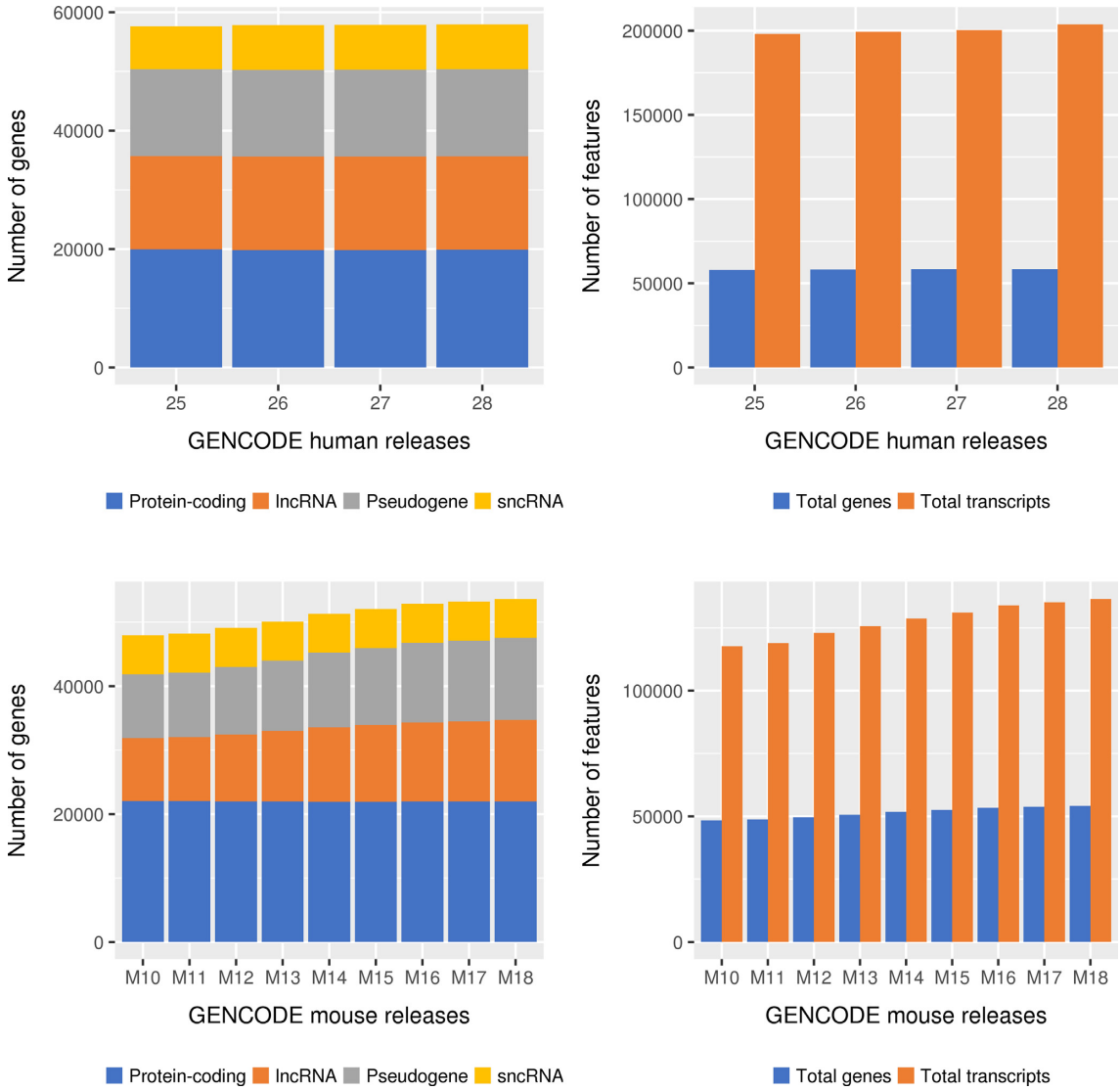

Figure 2. Annotation statistics for human and mouse GENCODE releases from July 2016 to June 2018, encompassing human releases GENCODE 25–28 and mouse releases M10 to M18. The panels on the left show the total number of genes by broad biotype (protein-coding, lncRNA, pseudogene and sncRNA) for each release for human and mouse respectively and panels on the right show the total numbers of genes and transcripts of all biotypes.

In mouse however, there is clear growth in the numbers of both genes and transcripts driven predominantly by the addition of lncRNAs and pseudogenes.

Extensive data resources for current and archival GENCODE releases are available at https://www.gencodegenes.org. As described above, the GENCODE gene sets are available as default in the Ensembl genome browser and also accessible via the UCSC genome browser. Other interfaces include the Ensembl FTP site (ftp://ftp.ensembl.org/pub/), which includes gene sets in GFF3, Genbank and GTF formats and full download of the complete Ensembl databases. More complex and customizable gene set queries can be created via the Ensembl Biomart (https://www.ensembl.org/biomart/).

Programmatic access to the GENCODE gene sets is possible via the extensive Ensembl Perl API and the language-agnostic Ensembl REST API. Programmatic access facilitates advanced genome-wide analysis such as retrieval of supporting features and associated gene trees. Examples of REST endpoint usage and starter scripts in different languages are at https://rest.ensembl.org.

GENCODE has been created exclusively on the GRCh38 human assembly since GENCODE 20 (August 2014). However, versions of selected releases since then that have been projection mapped from GRCh38 to GRCh37 are available at UCSC and from https://www.gencodegenes.org. Referred to as the 'lift37' annotation set, these data help identify genes where the annotations may have changed between GRCh37 and GRCh38. Due to the difficulty to generate accurate projections, the 'lift37' annotation set is not considered official reference annotation and only minimal support is available.

We welcome questions and feedback from the community directly via the helpdesks at https://www.gencodegenes.org, Ensembl and UCSC. In addition, the Ensembl and UCSC outreach activities annually reach thousands of re-

searchers via workshops at institutions and meetings, web-based training forums and 'how-to' guides focused on using the genome browsers and making best use of their features and data.

CONCLUSION

The GENCODE consortium continues to improve the quality of the reference gene annotation in human and mouse. We have integrated cutting-edge developments in the technology and scientific understanding of genome biology into our annotation workflows to improve the representation of existing loci and extend annotation coverage via the addition of entirely novel loci and alternatively spliced transcripts. While the high quality of our existing transcript annotation is extensively supported by both public data and data generated within the consortium, the abundance of evidence from new transcriptomic and proteomic datasets makes it clear that they are not yet complete.

ACKNOWLEDGEMENTS

We thank Tim Hubbard and Jennifer Harrow for their leadership in the GENCODE project from 2003-2016 as well as all groups and group members involved in the GENCODE project since its inception including the HAVANA manual annotation group formerly at Wellcome Sanger Institute now at EMBL-EBI (founder), the Guigo group at Centre for Genomic Regulation (founder), the Gerstein group at Yale (founder), the Center for Biomolecular Science & Engineering at UCSC (founder), the Ensembl team at EMBL-EBI (joined 2007), the Kellis group at MIT (joined 2007), the Tress group at CNIO (joined 2007), the Choudhary group formerly at Wellcome Sanger Institute now at Institute of Cancer Research (joined 2012), the Reymond group at University of Lausanne (2003–2017), the Antonarakis group at University of Geneva (2003–2007), the Wei group at Genome Institute of Singapore (2003–2007), the Gingeras group at Affymetrix Ltd (2003–2007) and the Brent group at Washington University in St. Louis (2007–2012).

FUNDING

National Human Genome Research Institute of the National Institutes of Health [U41HG007234]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health; Wellcome Trust [WT108749/Z/15/Z, WT200990/Z/16/Z]; European Molecular Biology Laboratory; Swiss National Science Foundation through the National Center of Competence in Research 'RNA & Disease' (to R.J.); Medical Faculty of the University of Bern (to R.J). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. Paul Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

REFERENCES

1. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
2. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
3. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.-K., Chrast,J., Lagarde,J., Gilbert,J.G.R., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**(Suppl. 1), doi:10.1186/gb-2006-7-s1-s4.
4. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
5. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
6. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
7. Casper,J., Zweig,A.S., Villarreal,C., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Karolchik,D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.
8. GTEx Consortium. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
9. International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
10. Stunnenberg,H.G., International,Human Epigenome Consortium and Hirst,M. (2016) The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, **167**, 1145–1149.
11. 1000, Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
12. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
13. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The Human Cell Atlas. *Elife*, **6**, e27041.
14. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
15. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
16. Pujar,S., O'Leary,N.A., Farrell,C.M., Loveland,J.E., Mudge,J.M., Wallin,C., Girón,C.G., Diekhans,M., Barnes,I., Bennett,R. *et al.* (2018) Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res.*, **46**, D221–D228.
17. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
18. Irimia,M., Weatheritt,R.J., Ellis,J.D., Parikshak,N.N., Gonatopoulos-Pournatzis,T., Babor,M., Quesnel-Vallières,M., Tapial,J., Raj,B., O'Hanlon,D. *et al.* (2014) A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, **159**, 1511–1523.
19. Tilgner,H., Jahanbani,F., Blauwkamp,T., Moshrefi,A., Jaeger,E., Chen,F., Harel,I., Bustamante,C.D., Rasmussen,M. and Snyder,M.P. (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.*, **33**, 736–742.
20. Abascal,F., Juan,D., Jungreis,I., Martinez,L., Rigau,M., Rodriguez,J.M., Vazquez,J. and Tress,M.L. (2018) Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res.*, **46**, 7070–7084.
21. Ezkurdia,I., Juan,D., Rodriguez,J.M., Frankish,A., Diekhans,M., Harrow,J., Vazquez,J., Valencia,A. and Tress,M.L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.*, **23**, 5866–5878.
22. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.

23. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
24. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)*, **2016**, baw093.
25. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
28. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
29. Hardwick,S.A., Chen,W.Y., Wong,T., Deveson,I.W., Blackburn,J., Andersen,S.B., Nielsen,L.K., Mattick,J.S. and Mercer,T.R. (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods*, **13**, 792–798.
30. Steijger,T., Abril,J.F., Engström,P.G., Kokocinski,F., Abril,J.F., Akerman,M., Alioto,T., Ambrosini,G., Antonarakis,S.E., Behr,J. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.
31. Lagarde,J., Uszczynska-Ratajczak,B., Carbonell,S., Pérez-Lluch,S., Abad,A., Davis,C., Gingeras,T.R., Frankish,A., Harrow,J., Guigo,R. *et al.* (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.*, **49**, 1731–1740.
32. Uszczynska-Ratajczak,B., Lagarde,J., Frankish,A., Guigó,R. and Johnson,R. (2018) Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.*, **19**, 535–548.
33. Weisser,H., Wright,J.C., Mudge,J.M., Gutenbrunner,P. and Choudhary,J.S. (2016) Flexible Data Analysis Pipeline for High-Confidence Proteogenomics. *J Proteome Res.*, **15**, 4686–4695.
34. Wright,J.C. and Choudhary,J.S. (2016) DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *J Proteomics Bioinform*, **9**, 176–180.
35. Wright,J.C., Mudge,J., Weisser,H., Barzine,M.P., Gonzalez,J.M., Brazma,A., Choudhary,J.S. and Harrow,J. (2016) Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.*, **7**, 11778.
36. Howald,C., Tanzer,A., Chrast,J., Kokocinski,F., Derrien,T., Walters,N., Gonzalez,J.M., Frankish,A., Aken,B.L., Hourlier,T. *et al.* (2012) Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.*, **22**, 1698–1710.
37. Fiddes,I.T., Armstrong,J., Diekhans,M., Nachtweide,S., Kronenberg,Z.N., Underwood,J.G., Gordon,D., Earl,D., Keane,T., Eichler,E.E. *et al.* (2018) Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res.*, **28**, 1029–1038.
38. Lilue,J., Doran,A.G., Fiddes,I.T., Abrudan,M., Armstrong,J., Bennett,R., Chow,W., Collins,J., Collins,S., Czechanski,A. *et al.* (2018) Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.*, doi:10.1038/s41588-018-0223-8.
39. Gordon,D., Huddleston,J., Chaisson,M.J.P., Hill,C.M., Kronenberg,Z.N., Munson,K.M., Malig,M., Raja,A., Fiddes,I., Hillier,L.W. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science*, **352**, aae0344.
40. Kronenberg,Z.N., Fiddes,I.T., Gordon,D., Murali,S., Cantsilieris,S., Meyerson,O.S., Underwood,J.G., Nelson,B.J., Chaisson,M.J.P., Dougherty,M.L. *et al.* (2018) High-resolution comparative analysis of great ape genomes. *Science*, **360**, eaar6343.
41. Paten,B., Earl,D., Nguyen,N., Diekhans,M., Zerbino,D. and Haussler,D. (2011) Cactus: algorithms for genome multiple sequence alignment. *Genome Res.*, **21**, 1512–1528.
42. König,S., Romoth,L.W., Gerischer,L. and Stanke,M. (2016) Simultaneous gene finding in multiple genomes. *Bioinformatics*, **32**, 3388–3395.
43. Rodriguez,J.M., Maietta,P., Ezkurdia,I., Pietrelli,A., Wesselink,J.-J., Lopez,G., Valencia,A. and Tress,M.L. (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, **41**, D110–D117.
44. Ezkurdia,I., Rodriguez,J.M., Carrillo-de Santa Pau,E., Vázquez,J., Valencia,A. and Tress,M.L. (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, **14**, 1880–1887.
45. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.*, **46**, D213–D217.
46. Zhang,Z., Carriero,N., Zheng,D., Karro,J., Harrison,P.M. and Gerstein,M. (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, **22**, 1437–1439.
47. Baertsch,R., Diekhans,M., Kent,W.J., Haussler,D. and Brosius,J. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, **9**, 466.
48. Navarro,F.C.P. and Galante,P.A.F. (2013) RCPedia: a database of retrocopied genes. *Bioinformatics*, **29**, 1235–1237.